"An empirical study about the use of Generative Adversarial Networks for text generation"

Trabajo Fin de Máster - Lenguajes y sistemas informáticos

Iván Vallés Pérez

Advisors:
Dr. D. Anselmo Peñas-Padilla
Dr. D. Emilio Soria-Olivas

28 de junio de 2018

Índice

Progressive GAN 2017 [6]

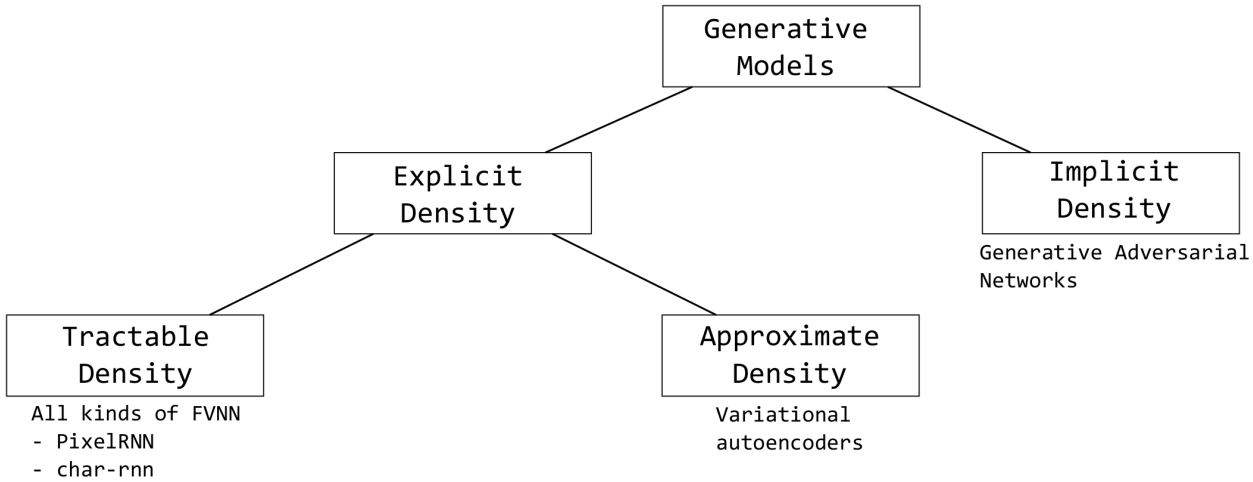
El objetivo principal del trabajo es estudiar la posibilidad de generar texto usando *Generative Adversarial Networks*

- Realizar un **estudio exhaustivo** sobre la **posibilidad de generar texto con GAN** usando algoritmos de optimización basados en el gradiente
- **Conseguir un modelo GAN estable (que no diverja)**. Estos algoritmos han mostrado dificultades[1] para generar datos de naturaleza discreta
- Diseñar una **arquitectura totalmente diferenciable** para poder aplicar **métodos de optimización basados en el gradiente**
- Entrenar el modelo de **generación de texto a nivel carácter**, de modo que la red pueda generar palabras no vistas en los datos de entrenamiento
- Lograr un **nivel de calidad alto** en el texto generado en los distintos aspectos del lenguaje: **morfológico**, **sintáctico**, **semántico**

[1] https://www.reddit.com/r/MachineLearning/comments/40ldq6/generative_adversarial_networks_for_text/cyyp0nl

Los modelos generativos aprenden de una distribución de datos real para generar nuevas muestras sintéticas

- Estas técnicas pertenecen al campo del **aprendizaje no supervisado** [1] ya que no necesitan datos etiquetados
- El proceso general consiste en **recopilar una gran cantidad de datos** de algún dominio (pueden ser de cualquier tipo: imágenes, texto, audio, video, datos estructurados, grafos, etc.) y entrenar a un modelo que generará nuevas muestras tratando de **imitar la distribución de los datos reales**



Existen tres familias básicas de modelos generativos basados en algoritmos de aprendizaje profundo

Modelos autoregresivos

Se basan en la **generación secuencial** de datos a partir de generaciones pasadas (p.e. generar una imagen píxel a píxel: PixelRNN [2])

Variational Autoencoders (VAE)

Consisten en un **autoencoder** (codificador-decodificador con cuello de botella) optimizado con la ayuda de métodos variacionales para que el código se ajuste a una distribución determinada [3]

Generative Adversarial Networks (GAN)

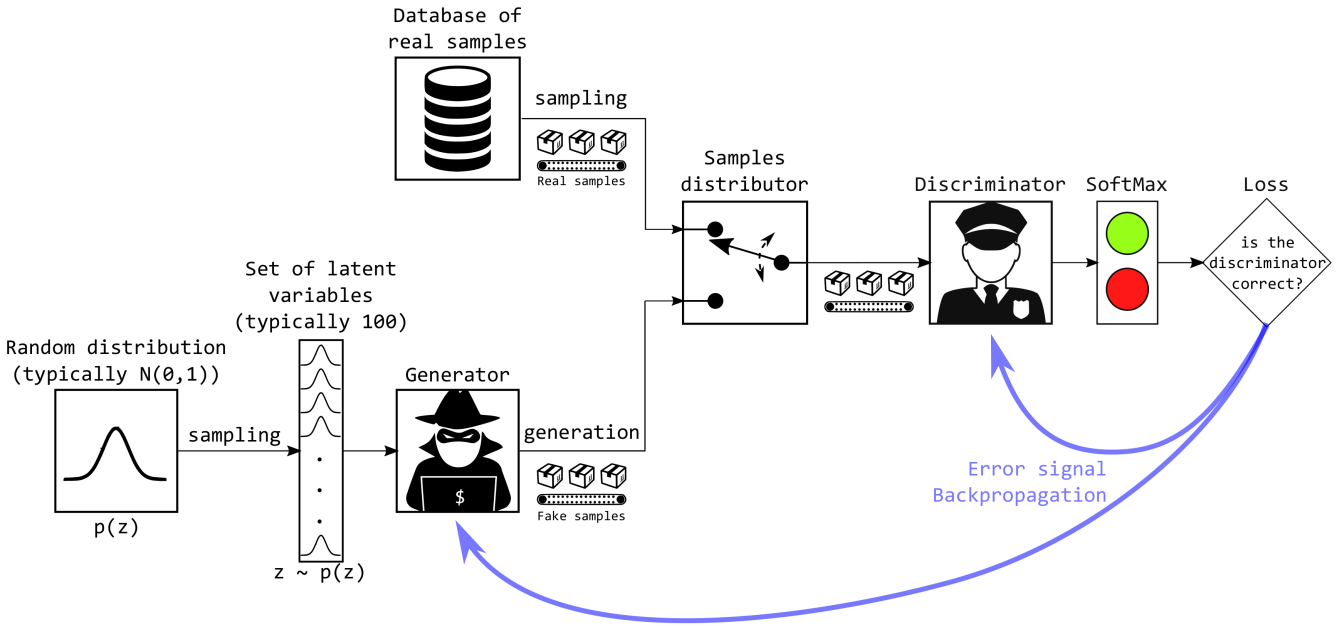
Constan de una red neuronal con **dos módulos**: un **generador** y un **discriminador** (llamado crítico en algunas versiones). El generador trata de generar muestras que se parezcan al conjunto de datos real mientras que el discriminador intenta distinguir entre muestras reales y generadas [4]

Las *Generative Adversarial Networks* son una familia de algoritmos nueva inventada por *Ian Goodfellow* en 2014

- Algoritmo inventado por *Ian Goodfellow* en 2014 [5]
- El objetivo del algoritmo es **aprender a representar una estimación de una distribución dada**
- Los modelos *GAN* son capaces de **aproximar la distribución de unos datos de entrada de forma implícita**, lo que permite tomar muestras de ellos, pero no se proporciona acceso a la distribución en sí
- El algoritmo original consiste en un **sistema de redes neuronales construido por 2 bloques principales: un generador y un discriminador** (también llamado crítico)
- Las *GAN* han demostrado ser **muy potentes en el campo de la visión computacional** [6]

Las *Generative Adversarial Networks* se componen de dos subredes: el generador y el discriminador

El modelo en conjunto puede entenderse como una **competición entre un falsificador** (el generador) **y un experto en falsificaciones** (el discriminador). El falsificador intenta imitar los datos reales lo mejor posible para que el experto no detecte las diferencias mientras que el experto intenta mejorar sus habilidades para cazar al falsificador.



WGAN-GP es una modificación de las *GAN* que consigue mejores resultados y mayor estabilidad

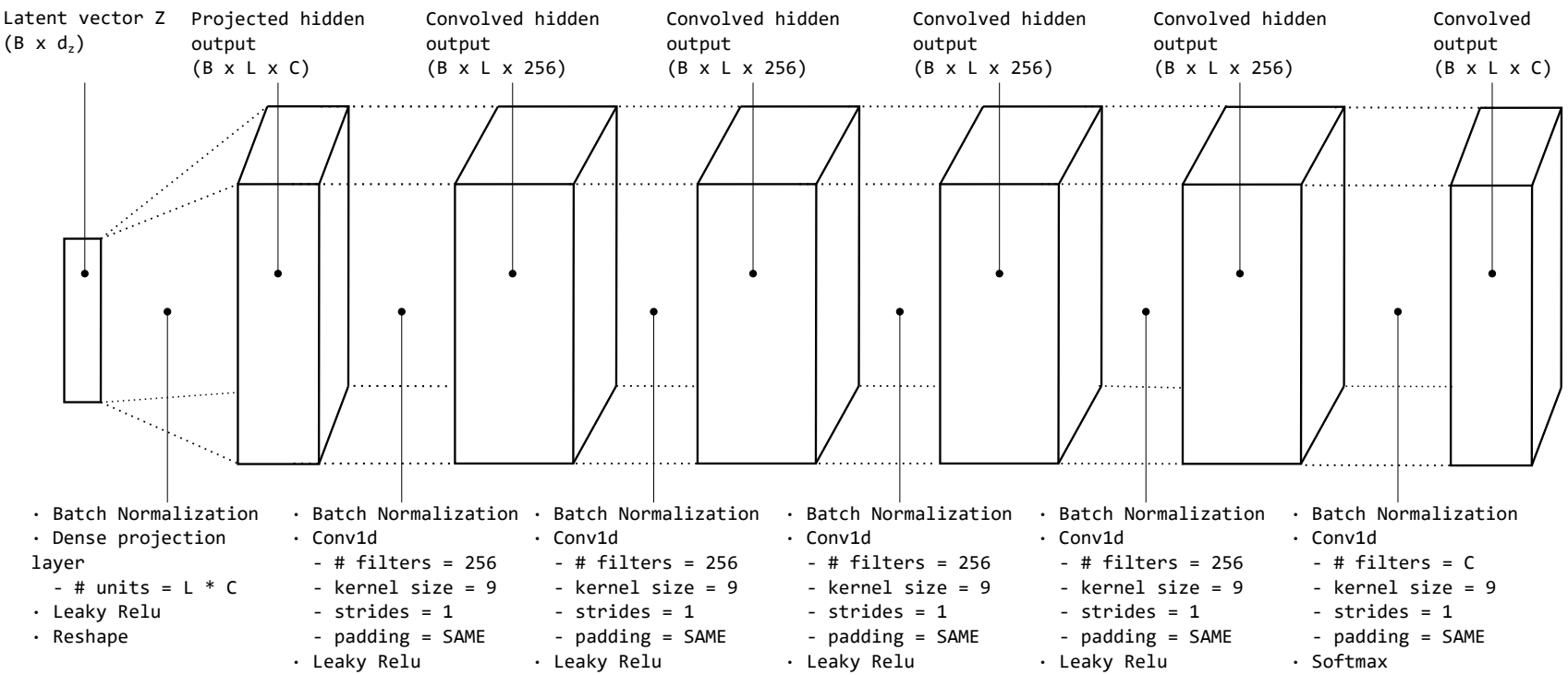
- **Los costes del generador y del discriminador de una GAN deben estar balanceados**. Si esta condición no se satisface, los gradientes del generador decrecen mucho o las actualizaciones de los parámetros de este se vuelven muy inestables [7] (el sistema diverge)
- **WGAN-GP** es una variación del algoritmo original que **soluciona parcialmente estos problemas** [8]
- Los modelos *WGAN-GP* minimizan la distancia de *Wasserstein-1*, la cual es continua y derivable en casi todo el espacio siempre que el sistema sea una función de *Lipschitz-1* [8]
- Para satisfacer la condicion *lipschitztiana*, los autores proponen añadir a la función de coste del discriminador un término de penalización de su gradiente

$$L_{critic} = \underbrace{\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)]}_{\text{original WGAN critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]}_{\text{gradient norm penalty}}$$

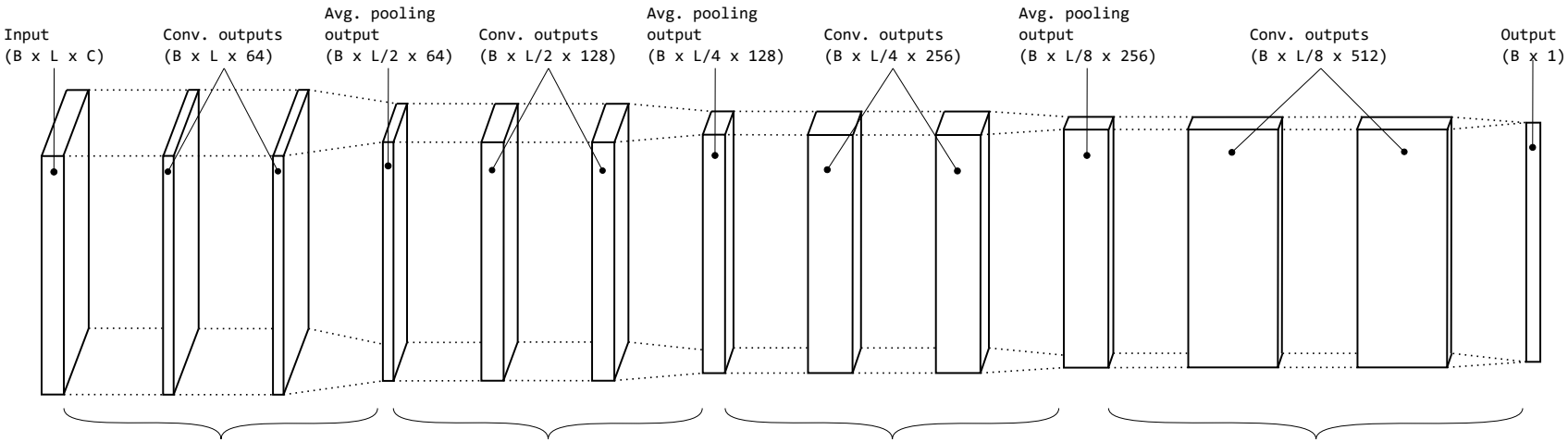
La mejor arquitectura de red neuronal obtenida se ha usado para generar los resultados reportados

- Se ha seguido un proceso empírico de **prueba y error para encontrar la arquitectura que mejores resultados produce** para este cometido
- Se han evaluado un total de **35 arquitecturas distintas**, de las cuales dos han funcionado sustancialmente mejor
- La arquitectura que **mejores resultados** ha mostrado consta de una **red neuronal convolucional** en el generador y otra en el discriminador
- El modelo final ha sido entrenado durante **65 días** en una *GPU Nvidia Titan X Pascal* para obtener el resultado presentado

El generador de la arquitectura propuesta contiene 6 bloques convolucionales con *batch normalization*



El discriminador de la arquitectura propuesta contiene 4 bloques convolucionales



Conv. block 1
- Conv1d + Conv1d
  - # filters = 64
  - kernel size = 9
  - strides = 1
  - padding = SAME
  - activation = leaky ReLU
- Average pooling
  - pool size = 2
  - strides = 2

Conv. block 2
- Conv1d + Conv1d
  - # filters = 128
  - kernel size = 9
  - strides = 1
  - padding = SAME
  - activation = leaky ReLU
- Average pooling
  - pool size = 2
  - strides = 2

Conv. block 3
- Conv1d + Conv1d
  - # filters = 256
  - kernel size = 9
  - strides = 1
  - padding = SAME
  - activation = leaky ReLU
- Average pooling
  - pool size = 2
  - strides = 2

Conv. block 4
- Conv1d + Conv1d
  - # filters = 512
  - kernel size = 9
  - strides = 1
  - padding = SAME
  - activation = leaky ReLU
- Average pooling
  - pool size = 8
  - strides = 2
- Conv1d
  - # Filters = kernel_size = strides = 1
  - padding = SAME
  - activation = None

Se ha usado la colección de datos *Tatoeba* para entrenar el modelo

- **Tatoeba**[2] **es una colección de oraciones escritas en múltiples idiomas**
- Se ha construido con ayuda de la **comunidad** y es **gratuita**
- Diseñada originalmente para tareas de traducción computacional
- Se han seleccionado sólo las **oraciones escritas en inglés** ($\sim 800K$)

```
Is it cruel to declaw your cat?
I knew that someone would come.
The eastern sky was getting light.
You are weak.
I very seldom eat lobster.
Those are the leftovers from lunch.
Do you like your new apartment?
It took him three tries.
Tom walked into his room.
```

[2] https://tatoeba.org

Los resultados han sido evaluados cuantitativamente y cualitativamente

Medición cuantitativa de resultados

- Se han diseñado una serie de métricas para medir cuantitativamente los resultados obtenidos por el modelo: **precisión k-gram**
- La métrica mide la **proporción de combinaciones de tuplas k-gram conocidas**, es decir, que han aparecido en el conjunto de datos real R.

$$Prec_k = \frac{\sum_{w \in S} g((w_1, ..., w_k))}{|S|} \quad \text{donde} \quad g((w_1, ..., w_k)) = \begin{cases} 1 & \text{if} \quad (w_1, ..., w_k) \in R \\ 0 & \text{if} \quad (w_1, ..., w_k) \notin R \end{cases}$$

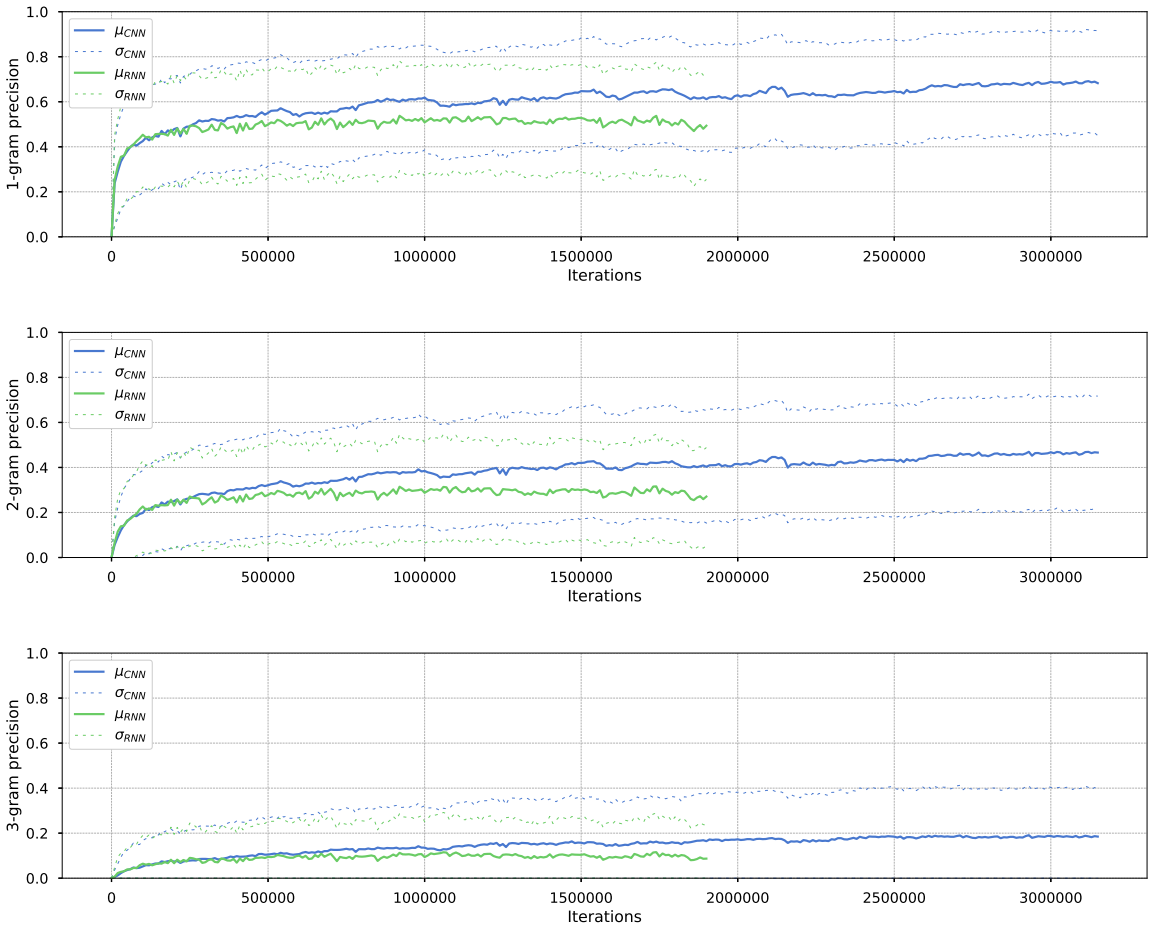
Medición cualitativa de resultados

- Dado que la **medición cuantitativa no es exhaustiva**, se ha realizado una **inspección manual de las oraciones generadas** para evaluar mejor la calidad de los resultados
- En estas inspecciones manuales **se han evaluado los aspectos morfológicos**, **sintácticos** y **semánticos** de las oraciones generadas

Los resultados muestran que el modelo ha sido capaz de aprender a escribir oraciones correctamente

- Desde el punto de vista **cuantitativo**, se han conseguido los siguientes resultados.
  - $\textbf{Precision}_{1-\textbf{gram}} = \textbf{67,3}\,\%$: el 67.3 % de las palabras generadas por el modelo son conocidas
  - $\textbf{Precision}_{2-\textbf{gram}} = \textbf{45,7}\,\%$: el 45.7 % de los bigramas generados por el modelo son conocidos
  - $\textbf{Precision}_{3-\textbf{gram}} = \textbf{18,6}\,\%$: el 18.6 % de los trigramas generados por el modelo son conocidos
- Una evaluación **cualitativa** concluye que el modelo es capaz de generar **muy buenos resultados a nivel morfológico**, **resultados notables a nivel sintáctico**, y **resultados aceptables a nivel semántico**.

Los gráficos de evolución de la precisión muestran que el modelo podría mejorar si se entrenase durante más tiempo



Las oraciones generadas muestran que el modelo ha sido capaz de generar texto a nivel carácter

Muestras con mayor calidad [3] obtenidas sobre una base de 5000 generaciones

```
I don't think Mary want to help Tom.
Why didn't you to talk Tom did that.
Tom didn't need to do that Tom.
What did Mary why they don't read.
What doesn't have to stay a job?
Tom said that you're still in this.
I did you do that for life.
Why the one you don't you.
Tom can't do that by him.
Tom and Mary said that here.
That has good to do that.
You didn't want to snow.
Where's not going to Tom?
It's just not to go.
```

```
I hope Tom would do that?
How many do that for me?
Tom said he wants to Boston.
Tom isn't do that alone.
You can't have her today.
How do you has to do.
Tom and  she doesn't have.
What did you that Tom did.
Did you do that the French?
Do you needed to do it?
I didn't had that before.
I know what you help him.
Tom didn't want to Boston.
Do you to like that.
```

[3] Calculada como el producto $Prec_1 \cdot Prec_2 \cdot Prec_3$. En el apéndice se adjunta una muestra de oraciones de baja calidad. En ambos casos las oraciones se han seleccionado sobre una base de 5000 generaciones. Se han eliminado las oraciones generadas que ya existen en los datos reales.

Las *GAN* son capaces de generar datos de naturaleza discreta con una calidad intermedia

- Se ha conseguido hallar una a**rquitectura totalmente diferenciable** que permite generar **texto a nivel carácter** mediante el uso de *GAN*
- El modelo conseguido es **estable** a la hora de **generar datos de naturaleza discreta**
- El texto generado por el mejor modelo obtenido (*arquitectura convolucional*) tiene una **calidad media**, que restringe la posibilidad de usar el modelo en entornos de producción
- Pese a haber sido capaz de generar texto, se concluye que las **GAN no deberían ser la mejor elección a la hora de abordar un problema de generación de texto complejo**

Referencias I

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[2] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA*, Jun. 2016.

[3] C. Doersch, "Tutorial on variational autoencoders," *Technical Report. Carnegie Mellon / UC Berkeley*, 2016.

[4] I. J. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *Neural Information Processing Systems*, Dec. 2016.

Referencias II

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *NIPS 2017 conference*, 2017.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (International Convention Centre, Sydney, Australia), pp. 214–223, PMLR, Aug. 2017.

Referencias III

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Backup

Las *Generative Adversarial Networks* constan de dos funciones de coste

- El discriminador toma x como datos de entrada y $\theta^{(D)}$ como parámetros. Consiste en una red que implementa un **clasificador binario** para determinar si la muestra de entrada x es real o generada.

$$J^{(D)}(\theta^{(D)}, \theta^{(G)}) = -\frac{1}{2} \cdot E_{x \sim p_{data}} (\log D(x)) - \frac{1}{2} \cdot E_z \log (1 - D(G(z)))$$

- El generador toma un vector z como datos de entrada y $\theta^{(G)}$ como parámetros. Esta red, provista de un vector z de variables aleatorias, se encarga de **generar muestras que intentan "engañar" al discriminador** para que las categorice como muestras reales.

$$J^{(G)}(\theta^{(D)}, \theta^{(G)}) = -E_z \log (D(G(z)))$$

$$\theta^{(D)} = \arg\min_{\theta^{(D)}} J^{(D)}(\theta^{(D)}, \theta^{(G)}) \qquad \theta^{(G)} = \arg\min_{\theta^{(G)}} J^{(G)}(\theta^{(D)}, \theta^{(G)})$$

No todas las oraciones generadas tienen calidad alta

Muestras con menor calidad[4] sobre una base de 5000 generaciones

```
I have thu back to eatentitive us.
Arnnot speal agread of the coping me.
When was everocly of same.
The thinks for you more indmistake.
Tom almays lidelh that.
She lookeded so from.
The racenstand lacking ourscomptay hease.
I had abbut sport I sddicisson.
Tom will got expreased?
Mary has onalis cold.
I call exput ite.
You should he've ever like him cheer.
Where can reat a loor truth?
Tom said Tom never importance.
```

```
I'll catrred me.
Ary you atselote.
Hen all the mongy rations.
Mryce the good oup if for.
I will sake.
I has just dright.
Tom has beemed to a red?
The trres.
He was pliting tbrue comprates.
Heat's oor her.
I readedinded him.
That's hia hacrtt it.
Who does here.
Tom wanted rightslea.
```

[4] Calculada como el producto $Prec_1 \cdot Prec_2 \cdot Prec_3$

No todas las oraciones generadas tienen calidad alta

Muestras escogidas aleatoriamente

He ever Tom and Mary what do that.
Mamy was father thank, bit Tom is in oriends.
The trribty didn't like early.
Tami wend by drrnished uso.
Iod make to fine heis catohing of cood.
He had to play on cletng from in today.
I'd mike a playing.
How long are I'm beree help us.
proli'd a callorate night is baik our.
My look is in the kinsered whire for leenlh.
It is a colurested and would awd from them.
You tharn I can book paraned.
They spaar Tom at nate ar friends.
Whas I want to stop over.

You much yourse.
I was works can.
Tom has now such.
I'm already.
What might hat been naws.
Dan Tom usked.
I'd fill me toam.
Mayle fonither ups.
He futter firud ons.
Don'd you help haply.
We'rery sice.
You are to ase saek.
I know that I caved.
I wouldn't know harried.